Technical Document 2277
May 1992

Misclassification Rates of Likelihood and Predictive Discriminant Functions for Small Samples

Don Waagen

Technical Document 2277
May 1992

Misclassification Rates of Likelihood and Predictive Discriminant Functions for Small Samples

Don Waagen

NAVAL COMMAND, CONTROL AND OCEAN SURVEILLANCE CENTER
RDT&E DIVISION
San Diego, California 92152-5000

J. D. FONTANA, C[illegible]T, USN
Commanding Officer

R. T. SHEARER
Executive Director

ADMINISTRATIVE INFORMATION

This work is funded by the National Security Agency, under program element 0305885G, project number CD38, and accession number ICCD38D0. The work was performed by Code 422 of the Naval Command, Control and Ocean Surveillance Center's RDT&E Division, San Diego, California.

Released by
Deborah Porter, Head
Intelligence Systems
Development Branch

Under authority of
J. A. Salzmann, Head
Ashore Command and
Intelligence Center Division

CONTENTS

FIGURES

CONTENTS (continued)

TABLES

CONTENTS (continued)

INTRODUCTION

In pattern recognition literature, and in the area of discriminant analysis, one commonly finds likelihood-based approaches to the classification problem of classifying a p-dimensional sample x into one class π_i from a population of classes π_i, ($i = 1, \ldots, m$), where the classes π_i are assumed to be multivariate normally distributed, with separate means μ_i and either a common covariance matrix Σ or individual covariance matrices Σ_i. Fisher's linear and quadratic discriminant functions are examples of likelihood-based discriminants, and are described and discussed in all pertinent literature. A good introduction is given by Duda and Hart [1973]. Geisser [1964] discusses a Bayesian approach to classification by deriving the predictive density for a class. In the Bayesian approach, classification of a new sample is based on the ratios of the predictive densities of the classes for the data sample x. The subject of this report is the comparison of the misclassification rates of predictive and likelihood discriminant functions under the constraint of small training sets and a varying number of training samples per class.

The literature comparing and contrasting the predictive and likelihood-based approaches is seemingly contradictory and definitely confusing. Kendall, Stuart, and Ord [1987] note that a complaint against the linear discriminant function is that it does not take into account the relative sizes of the training sets of the classes. They further state that the approach of predictive discrimination yields more reliable estimates than an equivalent likelihood approach. However, Raudys and Jain [1991] assert that Bayesian *density estimates (predictive densities) do not improve performance* over the quadratic discriminant function when sample sizes are different, and do not include the predictive discriminant function in their discussion of small training set classification.

Most literature acknowledges that the Fisher linear and quadratic discriminant functions are *asymptotically* optimal for Gaussian population classes (see Anderson [1984]). However, since optimality is asymptotic property, i.e., is true for large samples of the data, the functions are not necessarily optimal for small sample sizes. As a consequence, Enis and Geisser [1974] claim that the Bayesian-derived predictive density *is* optimal in minimizing the probability of misclassification. And asymptotically, the predictive density approaches the linear and quadratic in functional form. Thus it can be viewed that the optimality (as a function of sample size) of the likelihood-based discriminant functions is a function of how fast the predictive and likelihood densities converge.

In many fields a dilemma exists, where classification of data into a set of classes is desirable, but it is impossible or too expensive to obtain a reliable and large training set. Small training sets are therefore generally the rule in these fields. Classification is thus attempted by using a small number of training samples from each class. In addition, the number of training samples from each class is generally different. Small training samples and differing numbers of training samples for each class create a problem for the

likelihood-based discriminant functions. Marks and Dunn [1974] analyzed Fisher's linear, quadratic, and linear "best" discriminant functions, when sample size is small, i.e., $n = 10 - 100$. When sample sizes are moderate (i.e., $n = 100 - 500$), Wahl and Kronmal [1977] found that sample size was critical in choosing between the the linear and quadratic discriminant functions, even when the covariance matrices are unequal. However, both papers excluded an analysis of the properties of the predictive discriminant function, and this omission forms the basis for this report.

This report investigates the small-sample misclassification rates of traditional likelihood and predictive procedures. First, the likelihood and predictive approaches to classification are introduced. Second, Monte Carlo simulations compare the misclassification rates of the predictive density discriminator and Fisher's linear and quadratic discriminant functions. The misclassification rates of these functions are compared in the univariate case under the assumptions of the classes having (1) the same variance σ^2 and (2) differing variances σ_i^2. Simulations also vary parameters concerning class separation and sample size. These parameters are (1) the separation between class populations, (2) the number of training samples for each class, and (3) the total number of training samples. Third, a Monte Carlo simulation measures misclassification rates in a multivariate case using Fisher's Iris data. A conclusion follows the simulation results.

Note that the decision theory concept of associating a cost of misclassification with each class is not pursued in this report. For our analysis of the likelihood and predictive techniques of classification, the cost of misclassification is considered equal for all classes and is therefore not considered.

LIKELIHOOD APPROACH

Let the prior probabilities for classes π_1 and π_2 be given as p_1 and p_2, respectively. For a two-class case with known parameters, i.e., $\pi_1 \sim N(\mu_1, \Sigma_1)$, $\pi_2 \sim N(\mu_2, \Sigma_2)$, assign an observation $\mathbf{x}$ (a p-dimensional array) to a class in the following manner:

$$\text{Assign to class } \pi_1 : \quad \frac{L(\mathbf{x}|\mu_1, \Sigma_1)p_1}{L(\mathbf{x}|\mu_2, \Sigma_2)p_2} \geq 1$$

$$\text{Assign to class } \pi_2 : \quad \frac{L(\mathbf{x}|\mu_1, \Sigma_1)p_1}{L(\mathbf{x}|\mu_2, \Sigma_2)p_2} < 1$$

$L(\mathbf{x}|\mu_1, \Sigma_1)$ is the likelihood of observing $\mathbf{x}$ from a normal population $N(\mu_1, \Sigma_1)$. If the number of classes is greater than two, then the classification rule is modified to determine the most likely class given $\mathbf{x}$:

$$\text{Assign to class } \pi_k : \quad L(\mathbf{x}|\mu_k, \Sigma_k)p_k > L(\mathbf{x}|\mu_i, \Sigma_i)p_i \ \ \forall i \neq k$$

In the univariate case, we replace the covariance matrix Σ_i and mean vector μ_i with their respective scalar values σ_i^2 and μ_i :

$$\text{Assign to class } \pi_1 : \quad \frac{L(\mathbf{x}|\mu_1, \sigma_1^2)p_1}{L(\mathbf{x}|\mu_2, \sigma_2^2)p_2} \geq 1$$

$$\text{Assign to class } \pi_2 : \quad \frac{L(\mathbf{x}|\mu_1, \sigma_1^2)p_1}{L(\mathbf{x}|\mu_2, \sigma_2^2)p_2} < 1$$

The multiclass univariate rule is

$$\text{Assign to class } \pi_k : \quad L(\mathbf{x}|\mu_k, \sigma_k^2)p_k > L(\mathbf{x}|\mu_i, \sigma_i^2)p_i \ \ \forall i \neq k$$

When the parameters are unknown, one replaces the parameters with the "best" (in some sense) estimates for the parameters. In the multivariate normal case, we use for μ_1 and Σ_k

$$\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$$

$$(N_i - 1)S_i = \sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)$$

respectively, where N_i is the number of training samples and the s corresponds to training samples from class π_1.

The linear and quadratic discriminant functions will be derived for the two-class case with unknown parameters. Note, however, that the previously mentioned modification of the two-class case is easily applied to these functions to derive the multiple-class decision rules for the discriminant functions.

FISHER'S LINEAR DISCRIMINANT FUNCTION

If the covariance matrices (of dimension $p \times p$) are assumed equal but unknown, i.e., $\Sigma_1 = \Sigma_2 = \Sigma$, one can improve $\mathbf{S}$, the estimate of Σ, by using a weighted average of the sample variances:

$$(N_1 + N_2 - 2)\mathbf{S} = \sum_{i=1}^{N_1} (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1) + \sum_{j=1}^{N_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)$$

The likelihood decision rule becomes, with associated prior probabilities p_1 and p_2

$$\text{Assign to class } \pi_1 : \frac{(2\pi)^{1/2}|\mathbf{S}|^{1/2}\exp -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1)p_1}{(2\pi)^{1/2}|\mathbf{S}|^{1/2}\exp -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)p_1} \geq 1$$

otherwise assign to class π_2 (ratio is < 1). Cancelling the constants, multiplying the prior probabilities to the other side, and taking the logarithm of both sides, the equation can be written as the following:

$$\text{Assign to class } \pi_1 : \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{x}_2) - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) \geq \ln\frac{p_2}{p_1}$$

Expanding the left side and reducing, one is left with the linear discriminant function:

$$\text{Assign to class } \pi_1 : \mathbf{x}'\mathbf{S}^{-1}(\mathbf{x}_1-\bar{\mathbf{x}}_2) - \frac{1}{2}(\bar{\mathbf{x}}-\bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \geq \ln\frac{p_2}{p_1}$$

$$\text{Assign to class } \pi_2 : \mathbf{x}'\mathbf{S}^{-1}(\mathbf{x}_1-\bar{\mathbf{x}}_2) - \frac{1}{2}(\bar{\mathbf{x}}-\bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \geq \ln\frac{p_2}{p_1}$$

This equation describes a hyperplane decision surface, which is perpendicular to the vector between the means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$. For a derivation of these equations, see Anderson [1984].

For the univariate case, one replaces vectors with the corresponding scalar values. The linear discriminant function now becomes

$$\text{Assign to class } \pi_1 : \frac{x(\bar{x}_1-\bar{x}_2)}{S} - \frac{1}{2S}(\bar{x}_1+\bar{x}_2)(\bar{x}_1-\bar{x}_2) \geq \ln\frac{p_2}{p_1}$$

$$\text{Assign to class } \pi_2 : \frac{x(\bar{x}_1-\bar{x}_2)}{S} - \frac{1}{2S}(\bar{x}_1+\bar{x}_2)(\bar{x}_1-\bar{x}_2) \geq \ln\frac{p_2}{p_1}$$

If $p_1 = p_2$, then the logarithm of their ratio equals 0, and the univariate linear discriminant equation can be further reduced to the following:

$$\text{Assign to class } \pi_1 : x(\bar{x}_1 - \bar{x}_2) \geq \frac{1}{2}(\bar{x}_1 + \bar{x}_2)(\bar{x}_1 - \bar{x}_2)$$

$$\text{Assign to class } \pi_2 : x(\bar{x}_1 - \bar{x}_2) \geq \frac{1}{2}(\bar{x}_1 + \bar{x}_2)(\bar{x}_1 - \bar{x}_2)$$

This is equivalent to finding on which side observation x lies in relation to the midpoint of the line segment between the class means $\bar{x}_1$ and $\bar{x}_2$ (the hyperplane is a point). A graphical example of the univariate linear discriminant function for two classes, with $\bar{x}_1$ and $\bar{x}_2$ equal to 4.0 and 6.0, is given in figure 1.

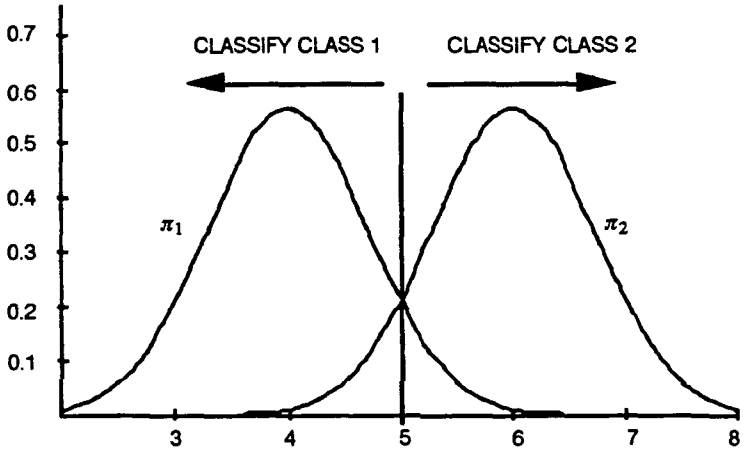

Figure 1. Linear discriminant function.

FISHER'S QUADRATIC DISCRIMINANT FUNCTION

In the p-dimensional multivariate case, if the covariance matrices of the classes are not considered equal, then the decision surface is not linear (a hyperplane), but quadratic. Hence the term quadratic discriminant function. Consider the case with two classes π_1 and π_2, with respective prior probabilities p_1 and p_2, sample mean vectors $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, and sample covariance matrices $\mathbf{S}_1^{-1}$ and $\mathbf{S}_2^{-1}$. The quadratic discriminant function is based on the ratio of the likelihood functions given x times the respective prior probability, and can be initially written (without simplification) as the following rule:

$$\text{Assign to class } \pi_1 : \frac{(2\pi)^{1/2}|\mathbf{S}_1|^{1/2}\exp -\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x} - \bar{\mathbf{x}}_1)p_1}{(2\pi)^{1/2}|\mathbf{S}_1|^{1/2}\exp -\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x} - \bar{\mathbf{x}}_1)p_2} \geq 1$$

Otherwise assign to class π_2.

This function can be written in a simpler form. Taking the logarithm of both sides and simplifying the equation gives a general form of quadratic discriminant function:

$$\text{Assign to class } \pi_1 : (\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) \leq \ln\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|} + 2\ln\frac{p_2}{p_2}$$

$$\text{Assign to class } \pi_2 : (\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) \leq \ln\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|} + 2\ln\frac{p_2}{p_2}$$

If the prior probabilities p_2 and p_1 are equal, a logarithm of their ratio equals 0, and the quadratic discriminant function can be rewritten as

$$\text{Assign to class } \pi_1 : (\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) \leq \ln\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}$$

$$\text{Assign to class } \pi_2 : (\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) \leq \ln\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}$$

For the univariate two-class case, one again replaces the vectors with scalar values, and the quadratic function can be written as

$$\text{Assign to class } \pi_1 : \frac{(x-\bar{x}_2)^2}{S_2} - \frac{(x-\bar{x}_1)^2}{S_1} \geq \ln\frac{S_2}{S_1}$$

$$\text{Assign to class } \pi_2 : \frac{(x-\bar{x}_2)^2}{S_2} - \frac{(x-\bar{x}_1)^2}{S_1} \geq \ln\frac{S_2}{S_1}$$

An example of the decision regions created by a univariate quadratic discriminant function is shown in figure 2.

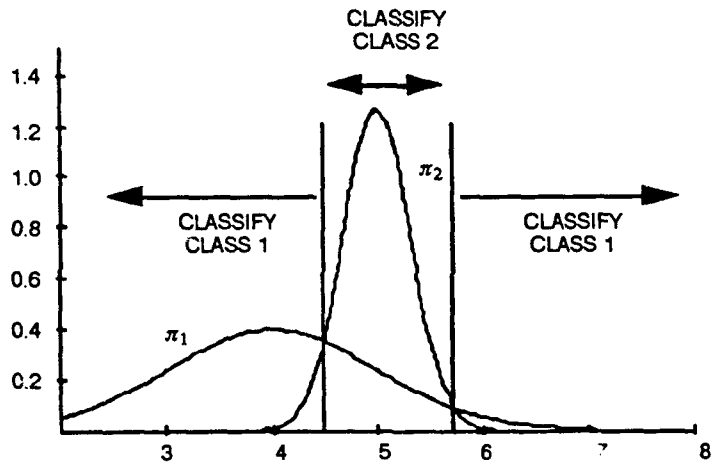

Figure 2. Quadratic univariate example.

THE PREDICTIVE DENSITY—A BAYESIAN APPROACH

A Bayesian approach to classification is based on computing the predictive density the probability $p(\mathbf{z} \mid \bar{\mathbf{x}}_i, \mathbf{S}_i, \pi_i)$, where $\mathbf{z}$ is a future (p-dimensional) observation. The predictive density replaces the likelihood function in the classification rule. The predictive density incorporates knowledge about how many sample points are being used to estimate the density, and therefore the predictive density becomes a function of the sample size N_i. Given definitions of $\bar{\mathbf{x}}_i$, $\mathbf{S}_i$, based on N_i observations previously defined, a Bayesian derivation of the probability that the observation $\mathbf{z}$ belongs to class π_i (i.e., $p(\pi_i \mid \bar{\mathbf{x}}_i, \mathbf{S}_i, \mathbf{z})$ for $i = 1, \ldots, K$) follows. For details on the Bayesian approach, see Press [1982]. The following derivation is due to Geisser [1964].

Define the joint prior probability density of π_i and Σ_i^{-1} as the following:

$$g(\Sigma_i)\sigma\pi_i\sigma\Sigma_i^{-1} \propto |\Sigma_i|^{(p+1)/2}\sigma\pi_i\sigma\Sigma_i^{-1}$$

where p is the dimension of μ_i. This joint prior probability is called a reference prior and indicates no prior knowledge of the distribution. Another constraint is that $p < N_i$, so the covariance matrix is not singular. Now we have from Geisser [1964, p.72] that the joint density of $\bar{\mathbf{x}}_i$, $\mathbf{S}_i$, conditional on the parameters μ_i, Σ_i^{-1}, and π_i, has the following form:

$$p(\bar{\mathbf{x}}_i, \mathbf{S}_I|\mu_i, \Sigma_i^{-1}, \pi_i) \propto |\mathbf{S}_i|^{(N_i-2-p)/2}|\Sigma_i|^{-N_i/2} \times \exp\{-\tfrac{1}{2}\mathrm{tr}\Sigma_i^{-1}[(N_i-1)\mathbf{S}_i + (\bar{\mathbf{x}}_i-\mu_i)(\bar{\mathbf{x}}_i-\mu_i)']\}$$

Multiplying $p(\bar{\mathbf{x}}_i, \mathbf{S}_i \mid \mu_i, \Sigma_i^{-1}, \pi_i)$ by the joint prior distribution gives

$$p(\mu_i, \Sigma_i^{-1} \mid \bar{\mathbf{x}}_i, \mathbf{S}_i, \pi_i) \propto |\Sigma_i|^{(N_i-p-1)/2}\exp\{-\tfrac{1}{2}\mathrm{tr}\Sigma_i^{-1}[(N_i-1)\mathbf{S}_i + (\bar{\mathbf{x}}_i-\mu_i)(\bar{\mathbf{x}}_i-\mu_i)']\}$$

Integrating over the parameters μ_i and Σ_i^{-1}, the predictive density for an observation $\mathbf{z}$ is obtained:

$$p(\mathbf{z} \mid \bar{\mathbf{x}}_i, \mathbf{S}_i, \pi_i) = \int\int p(\mathbf{x} \mid \mu_i, \Sigma_i^{-1}, \pi_i)p(\mu_i, \Sigma_i^{-1}|\bar{\mathbf{x}}_i, \mathbf{S}_i, \pi_i)\sigma\mu_i\sigma\Sigma_i^{-1}$$

$$= \left(\frac{N_i}{(N_i+1)\pi}\right)^{1/2p}\left\{\frac{\Gamma\left(\frac{N_i}{2}\right)}{\Gamma\left(\frac{N_i-p}{2}\right)|(N_i-1)\mathbf{S}_i|^{1/2}}\right\}\left(1+\frac{N_i}{N_i^2+1}(\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)\right)^{1/2N_i}$$

Classification of an observation $\mathbf{z}$ into a class π_i is according to

$$p(\pi_i \mid \mathbf{z}, \bar{\mathbf{x}}_i, \mathbf{S}_i) \propto p(\mathbf{z} \mid \bar{\mathbf{x}}_i, \mathbf{S}_i, \pi_i)p_i$$

where p_i is the prior probability of class π_i. Therefore, the two-class decision rule is the following:

$$\text{Assign to class } \pi_1 : \frac{p(\pi_1 \mid \mathbf{z}, \bar{\mathbf{x}}_i, \mathbf{S}_1)}{p(\pi_2 \mid \mathbf{z}, \bar{\mathbf{x}}_2, \mathbf{S}_2} \geq 1$$

$$\text{Assign to class } \pi_2 : \frac{p(\pi_1 \mid \mathbf{z}, \bar{\mathbf{x}}_i, \mathbf{S}_1)}{p(\pi_2 \mid \mathbf{z}, \bar{\mathbf{x}}_2, \mathbf{S}_2} < 1$$

This is the *predictive-odds ratio* for classifying $\mathbf{z}$ into π_1 as compared with π_2. Expanding the decision rule gives the following result:

Assign to class π_1 if

$$\frac{\left(\frac{N_1}{(N_1+1)\pi}\right)^{1/2p}\left\{\frac{\Gamma\left(\frac{N_1}{2}\right)}{\Gamma\left(\frac{N_1-p}{2}\right)|(N_1-1)\mathbf{S}_1|^{1/2}}\right\}\left(1+\frac{N_1}{N_1^2-1}(\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1)\right)^{1/2N_1}p_1}{\left(\frac{N_2}{(N_2+1)\pi}\right)^{1/2p}\left\{\frac{\Gamma\left(\frac{N_2}{2}\right)}{\Gamma\left(\frac{N_2-p}{2}\right)|(N_2-1)\mathbf{S}_2|^{1/2}}\right\}\left(1+\frac{N_2}{N_2^2-1}(\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)\right)^{1/2N_2}p_2} \geq 1$$

Otherwise assign the observation to class π_2. This can be simplified to the following form:

$$\text{Assign to class } \pi_1 : \quad K_{12}\frac{\left\{1+\frac{N_2}{N_2^2+1}(\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)\right\}^{1/2N_2}}{\left\{1+\frac{N_1}{N_1^2-1}(\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1)\right\}^{1/2N_1}} \geq 1$$

otherwise class π_2, where K_{12} is a constant not depending on $\mathbf{x}$

$$K_{12} = \left(\frac{p_1}{p_2}\right) \left(\frac{N_1(N_2+1}{N_2(N_1+1)}\right)^{1/2p} \left\{ \frac{\Gamma\left(\frac{N_1}{2}\right)\Gamma\left(\frac{N_2-p}{2}\right)}{\Gamma\left(\frac{N_2}{2}\right)\Gamma\left(\frac{N_1-p}{2}\right)} \right\} \left(\frac{|(N_2-1)\mathbf{S}_2|}{|(N_1-1)\mathbf{S}_1|}\right)^{1/2}$$

The predictive discriminant function is the ratio of two multivariate Student t-distributions. This gives the function several nice properties. One, the sample sizes need not be the same for the function to be applicable, as is implicit in the likelihood-based approaches. Two, since the discriminant function is a quadratic function, the covariance matrices need not be equal, which is required (or assumed) by Fisher's linear discriminant function. Figure 3 is an example of the predictive distributions generated from two t-distributions with equal sample variances but unequal number of samples per class.

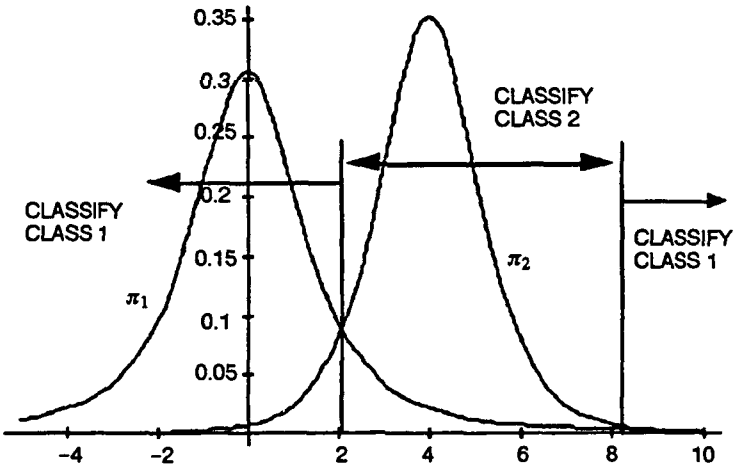

Figure 3. Univariate predictive distributions for $N(0,1)$ with N=3 vs. $N(4,1)$ with N=6.

The extension of the two-class case to a multiclass decision rule is simply

$$\text{Assign to class } \pi_k : \quad p(\pi_k|\mathbf{z}, \bar{\mathbf{x}}_k, \mathbf{S}_k) > p(\pi_2|\mathbf{z}, \bar{\mathbf{x}}_i, \mathbf{S}_i) \ \forall i \neq k$$

For the univariate case, the decision rule is simplified to the following:

$$\text{Assign to class } \pi_1: \quad K_{12} \frac{\left\{1+\frac{N_2}{N_2^2+1}\frac{(x-\bar{x}_2)^2}{S_2}\right\}^{1/2N_2}}{\left\{1+\frac{N_1}{N_1^2+1}\frac{(x-\bar{x}_1)^2}{S_1}\right\}^{1/2N_2}} \geq 1$$

where K_{12} is defined as

$$K_{12} = \left(\frac{p_1}{p_2}\right)\left(\frac{N_1(N_2+1}{N_2(N_1+1)}\right)^{1/2p}\left(\frac{\Gamma\left(\frac{N_1}{2}\right)\Gamma\left(\frac{N_2-p}{2}\right)}{\Gamma\left(\frac{N_2}{2}\right)\Gamma\left(\frac{N_1-p}{2}\right)}\right)\left(\frac{|(N_2-1)\mathbf{S}_2|}{|(N_1-1)\mathbf{S}_1|}\right)^{1/2}$$

The multivariate (univariate) t-distribution incorporates the information of the variability of the mean and variance attributable to the number of training samples. It is asymptotically normal, and therefore one could view the linear and quadratic functions as asymptotically approaching a t-distribution as n gets large (rather than the converse). Note that the number of points needed for the multivariate likelihood and predictive discriminant functions is a function of the dimensionality of the data. The higher the dimensionality, a greater number of training samples are needed for a reasonable and reliable estimation of the parameters. The relationship of class training sample size to misclassification rates, and therefore discriminant performance, will be investigated in the next section.

NUMERICAL SIMULATION AND ANALYSIS

To understand the relationship and capabilities of the discriminant functions discussed in the previous sections, Monte Carlo simulations have been made to measure the effects of modifying (1) the parameter values of the underlying class populations, and (2) the number of training samples for each class, for the misclassification rates of the discriminant functions. A Monte Carlo analysis of the relationship between the training sample size and misclassification rates of the discriminant functions is also performed on a multivariate data set, i.e., Fisher's Iris data.

UNIVARIATE TWO-CLASS CASE

The analysis of the univariate two-class case was performed as follows: The experimental parameters of population class variance and training sample size were set at various values, and the probabilities of misclassification were derived through Monte Carlo simulation. The simulations were performed according to the following algorithm:

1. For each class, the population parameter values, and the number of training samples, were sent via the argument list to the program.

2. The program generated a training set (of the specified number) as well as a test set of random numbers (currently 1000). Half of these test cases were generated from a normal distribution with class π_1 specified parameters (i.e., mean and variance). The other half of the test cases were generated from a normal distribution with class π_2 specified parameters.

3. Each test point (from both classes) was classified to belong to class π_1 or class π_2 by the discriminant functions. A count of the misclassifications was kept and the results were printed.

4. Steps 2 and 3 were repeated a number of times (300 in the cases presented below), generating misclassification samples for the discriminant functions, as a function of the parameter values passed to the program. The average of these misclassification samples was output as the result.

MISCLASSIFICATION RATES AS A FUNCTION OF TRAINING SIZE AND RATIO OF VARIANCES

To compare the capabilities of the three discriminant functions previously discussed in the small training sample problem, the average misclassification rates of the discriminant functions were measured and compared over several different training sample cases. With the different training sample sizes, and for each case of training sample sizes, the

variance of class π_2 ranged in value. For the given class training sample, an analysis was performed of the effect of the relative variance ratio between the classes and discriminant misclassification rate. In all cases, the prior probabilities of the two classes were considered equal. The details of the analysis follows.

Class π_1 was defined to have a normal (0,1) distribution. Class π_2 had a normal $(4, \sigma^2)$ distribution, where σ^2 was varied from $\frac{1}{256}$ to 256. Figure 4 graphically details several examples of the class π_1 and π_2 configurations. Since the "distance" between two univariate populations is, in the Mahalanobis sense, a function of the variances of the populations, the distance between the classes was modified despite the fact that the centers of the distributions were kept constant throughout the analysis. By allowing the variance of class π_2 to vary in value, the distance between the two classes can be viewed as a function of the class π_2 variance.

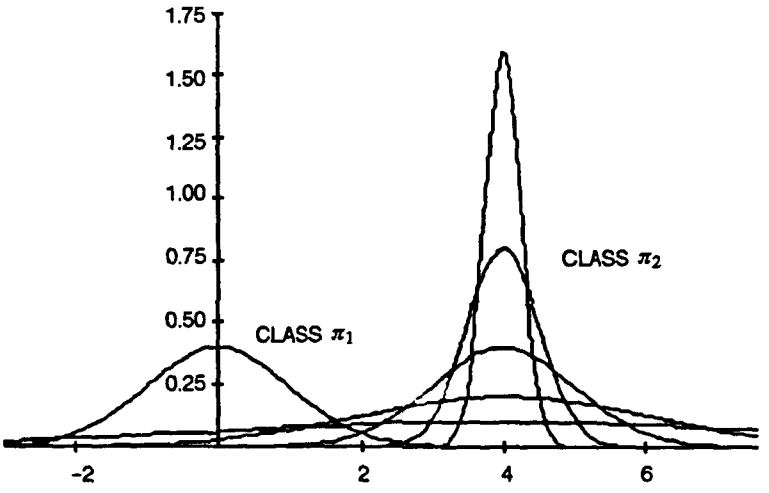

Figure 4. Class π_1 and class π_2 populations, showing several π_2 distributions with different variances.

The simulations were performed for several cases of equal and unequal training samples (figures 5–14 and tables 1–10). In case 1, both classes had an equal number of training samples. Case 2 was an example of unequal samples from the classes. Case 3 reversed the unequal number of samples from each class, to see if the discriminant functions were biased by sample size. Cases 4, 5, and 6 provided more information about various sample size configurations and the average misclassification rates for the discriminant functions. Cases 7 and 8 investigated the misclassification rates when distance between the means of the classes was increased and decreased, respectively. Cases 9 and 10 explored the variation of misclassification rates caused by varying the sample size of one of the distributions.

An explanation of figures 5-14 is in order at this time. The range of the plot is a logarithmic scale, based on the ratio of the standard deviations of the class populations. This linearizes the ratio of the standard deviation, moving the case of equal standard deviations to the point 0, and assigns equidistant points to reciprocal values of the ratio.

Case 1. Equal number of training samples for each class.

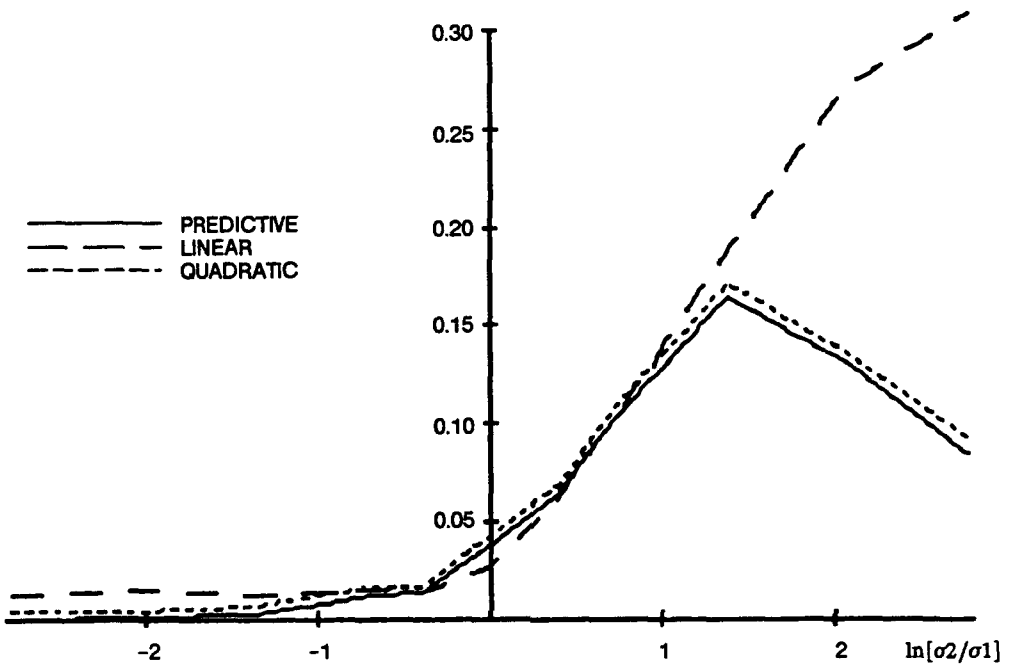

Figure 5. Probability of misclassification for various class π_2 variances, given an equal number of training samples $N_1 = N_2 = 6$.

Table 1. Average misclassification rates for discriminant functions, given various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = N_2 = 6$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0126	0.0126	0.0138	0.0143	0.0278	0.0628	0.0989	0.1876	0.3078
quad	0.0031	0.0059	0.0164	0.0165	0.0429	0.0695	0.1053	0.1712	0.0921
pred	0.0002	0.0025	0.0139	0.0122	0.0381	0.0647	0.0998	0.1644	0.0843

Case 2. Unequal number of training samples for each class.

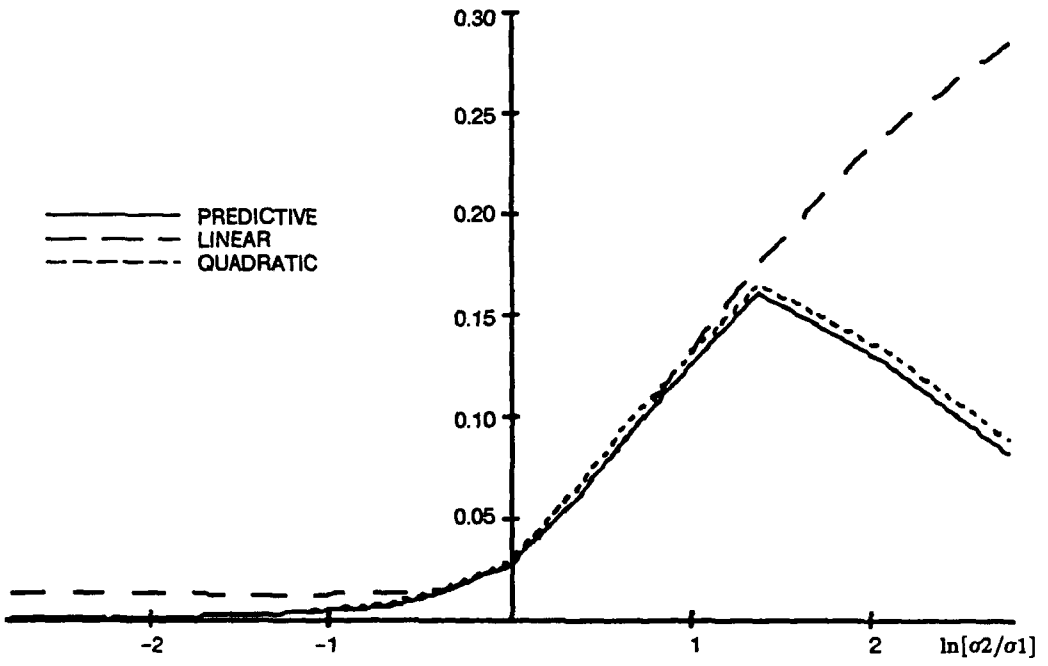

Figure 6. Probability of misclassification for various class π_2 variances, given an unequal number of training samples $N_1 = 6$, $N_2 = 18$.

Table 2. Misclassification rates for discriminant functions, given various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = 6$, $N_2 = 18$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0123	0.0119	0.0124	0.0148	0.0259	0.0610	0.0939	0.1750	0.2857
quad	0.0003	0.0019	0.0077	0.0148	0.0297	0.0671	0.1012	0.1646	0.0889
pred	0.0001	0.0014	0.0061	0.0125	0.0279	0.0622	0.0945	0.1603	0.0814

These results are fairly comparable to case 1, with the increased number of training samples from class 2 lowering the misclassification rates of all the discriminant functions.

Case 3. Unequal number of training samples—numbers exchanged from case 2.

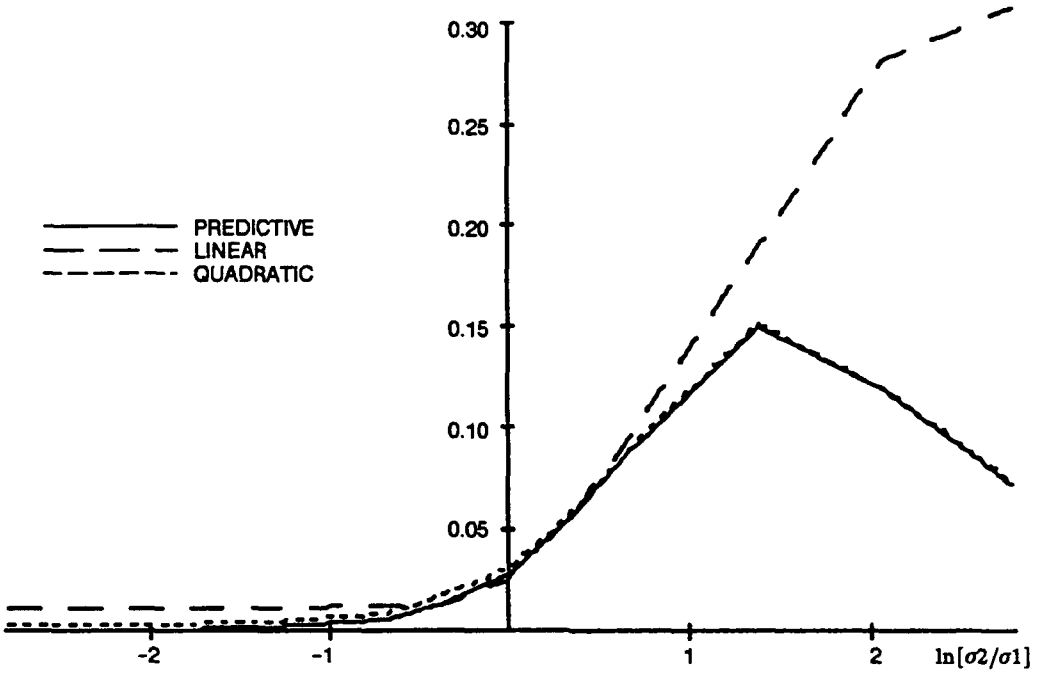

Figure 7. Probability of misclassification for various class π_2 variances, given an unequal number of training samples $N_1 = 18$, $N_2 = 6$.

Table 3. Misclassification rates for discriminant functions, given various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = 18$, $N_2 = 6$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0118	0.0118	0.0121	0.0128	0.0254	0.0621	0.0981	0.1899	0.3092
quad	0.0030	0.0049	0.0088	0.0170	0.0315	0.0612	0.0926	0.15143	0.0736
pred	0.0002	0.0017	0.0064	0.0127	0.0281	0.0599	0.0917	0.1499	0.0726

The results are comparable to case 2. Note that the quadratic function misclassification rate is an order of magnitude larger than its value in case 2 when the ratio of standard deviations is $\frac{1}{16}$.

Case 4. Differing number of training samples.

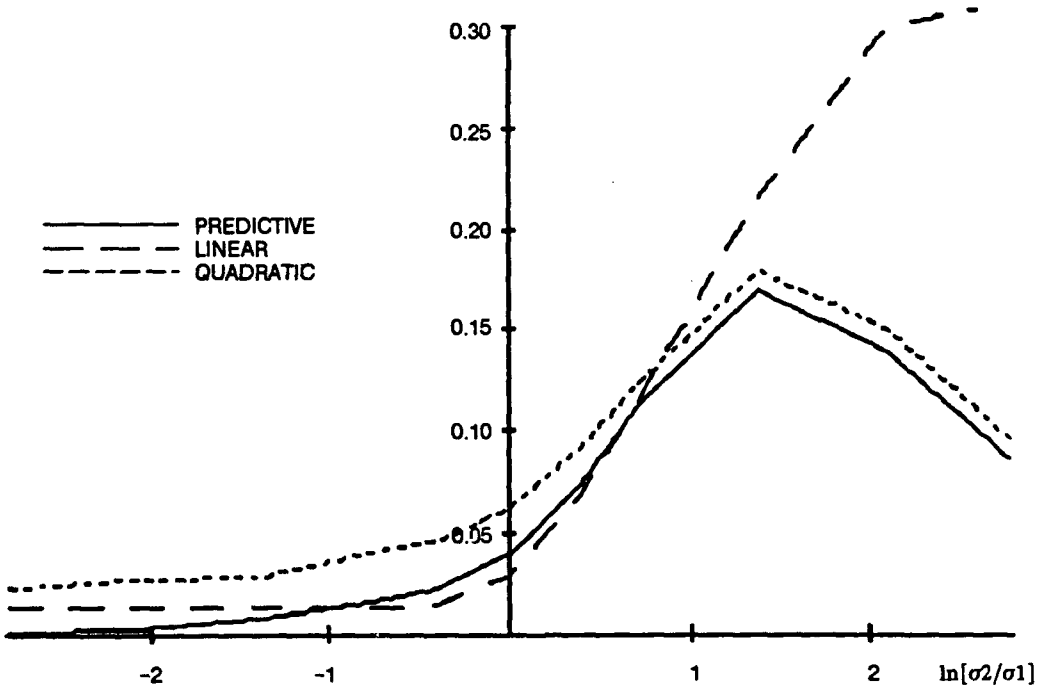

Figure 8. Probability of misclassification for various class π_2 variances, given an unequal number of training samples $N_1 = 6$, $N_2 = 3$.

Table 4. Average misclassification rates for discriminant functions for various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = 6$, $N_2 = 3$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0141	0.0146	0.0146	0.0149	0.0296	0.0685	0.1101	0.2151	0.3162
quad	0.0227	0.0279	0.0416	0.0460	0.0625	0.0916	0.1211	0.1786	0.0958
pred	0.0013	0.0075	0.0164	0.0228	0.0400	0.0746	0.1082	0.1697	0.0871

Case 5. Exchanged number of training samples from case 4.

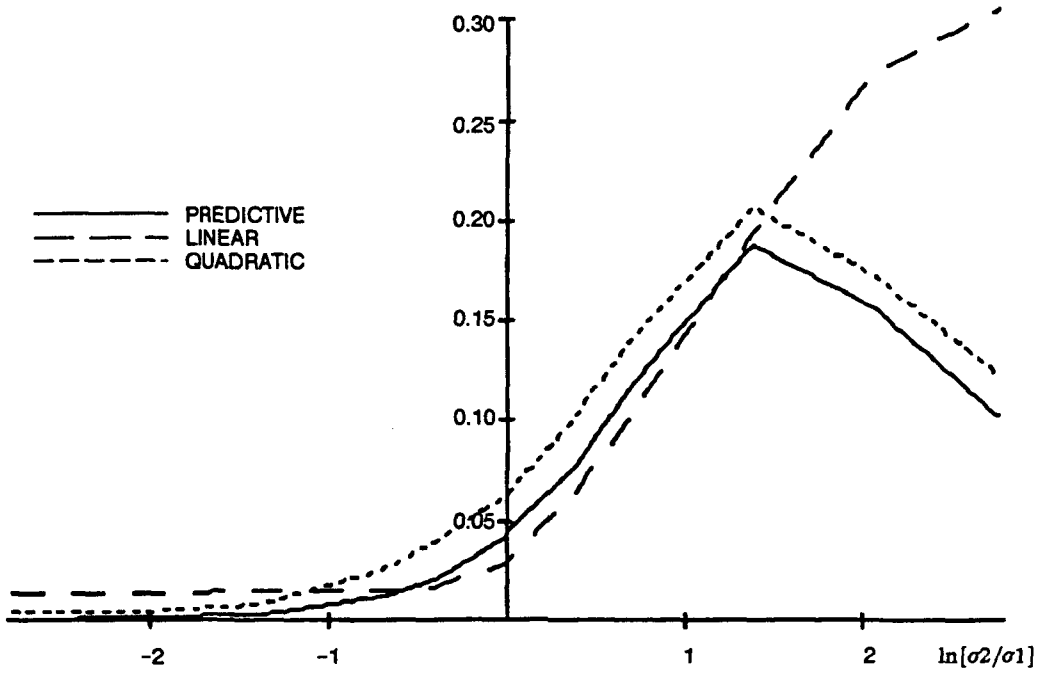

Figure 9. Probability of misclassification for various class π_2 variances, given an unequal number of training samples $N_1 = 3$, $N_2 = 6$.

Table 5. Average misclassification rates for discriminant functions for various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = 3$, $N_2 = 6$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0141	0.0146	0.0146	0.0294	0.0294	0.0659	0.1018	0.1952	0.3051
quad	0.0042	0.0088	0.0256	0.0394	0.0629	0.1043	0.1407	0.2062	0.1228
pred	0.0004	0.0037	0.0128	0.0210	0.0437	0.0795	0.1164	0.1877	0.1033

Case 6. Extremely small number size for class π_2.

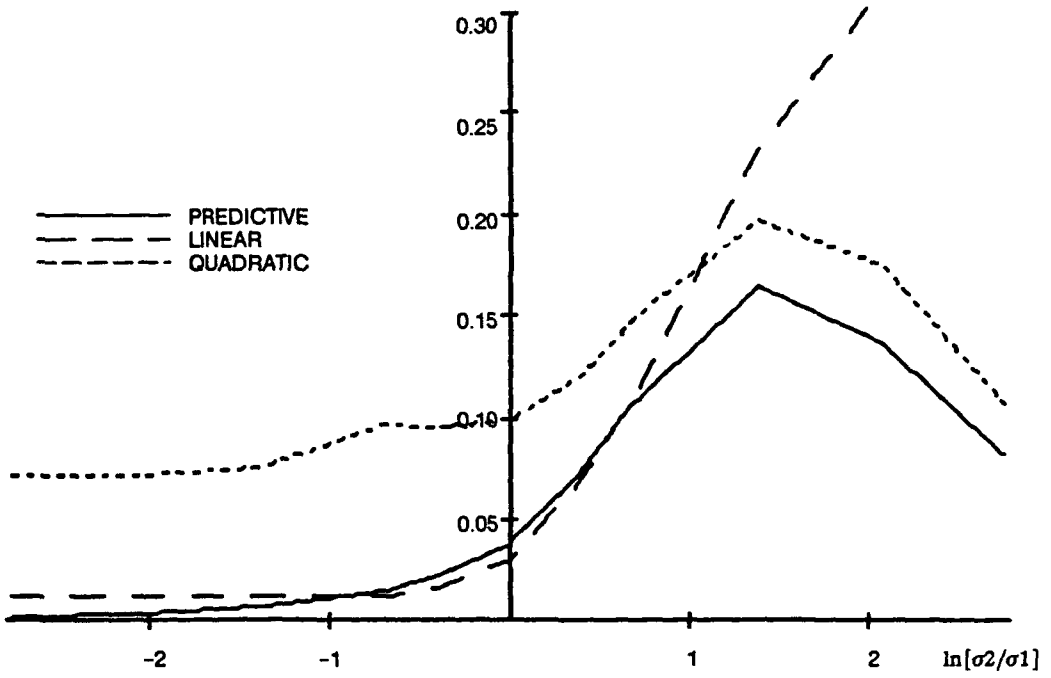

Figure 10. Probability of misclassification for various class π_2 variances, given an unequal number of training samples $N_1 = 10$, $N_2 = 2$.

Table 6. Average misclassification rates for discriminant functions for various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = 10$, $N_2 = 2$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0117	0.0116	0.0128	0.0155	0.0306	0.0695	0.1102	0.2320	0.3183
quad	0.0718	0.0765	0.0962	0.0956	0.0989	0.1197	0.1494	0.1951	0.1070
pred	0.0023	0.0076	0.0143	0.0224	0.0384	0.0739	0.1088	0.1651	0.0827

The behavior of the quadratic discriminant function in case 4 and the above case clearly shows an inability to handle small sample sizes even when the distributions are quite separated. Indeed, the quadratic performance is worse than the linear in these cases, even when the variances are quite different. The predictive distribution does not exhibit

the same problem, but handles small sample sizes well across all values of the class π_2 standard deviations.

Case 7. Distance between means of classes is increased.

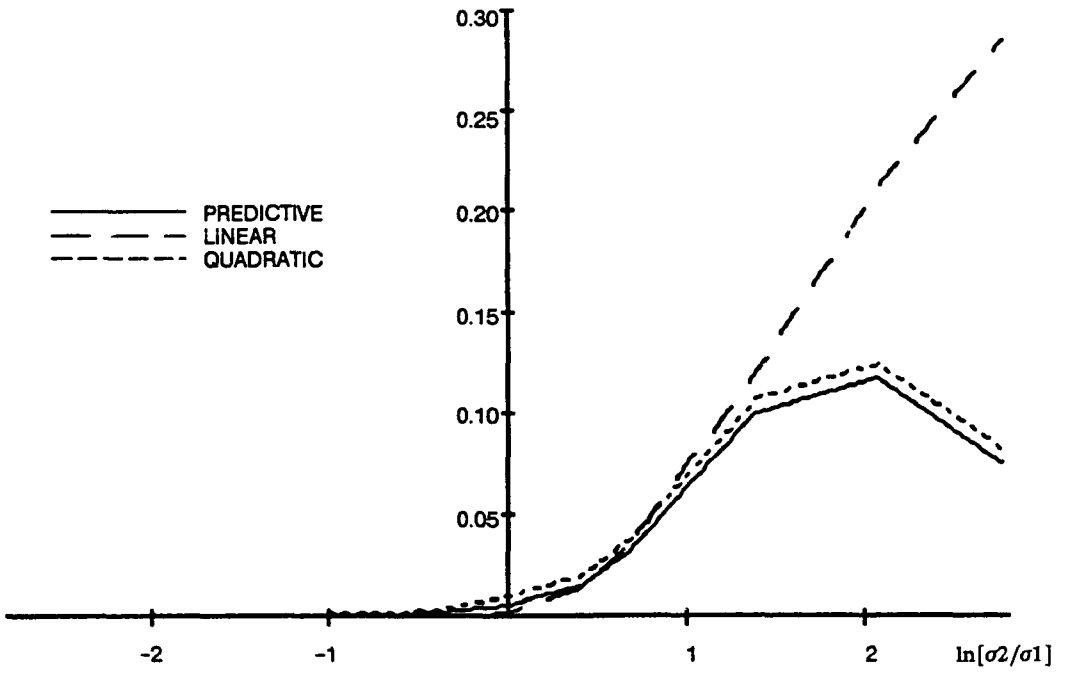

Figure 11. Average misclassification rates when distance between means is increased $N_1 = N_2 = 6$.

Table 7. Average misclassification rates when distance between means is increased, for various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = N_2 = 6$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0008	0.0008	0.0008	0.0009	0.0019	0.0143	0.0375	0.1189	0.2853
quad	0.0004	0.0006	0.0020	0.0032	0.0094	0.0198	0.0386	0.1062	0.0828
pred	0.0000	0.0001	0.0009	0.0018	0.0064	0.0160	0.0330	0.1002	0.0757

Case 8. Distance between means is decreased, with equal sample sizes.

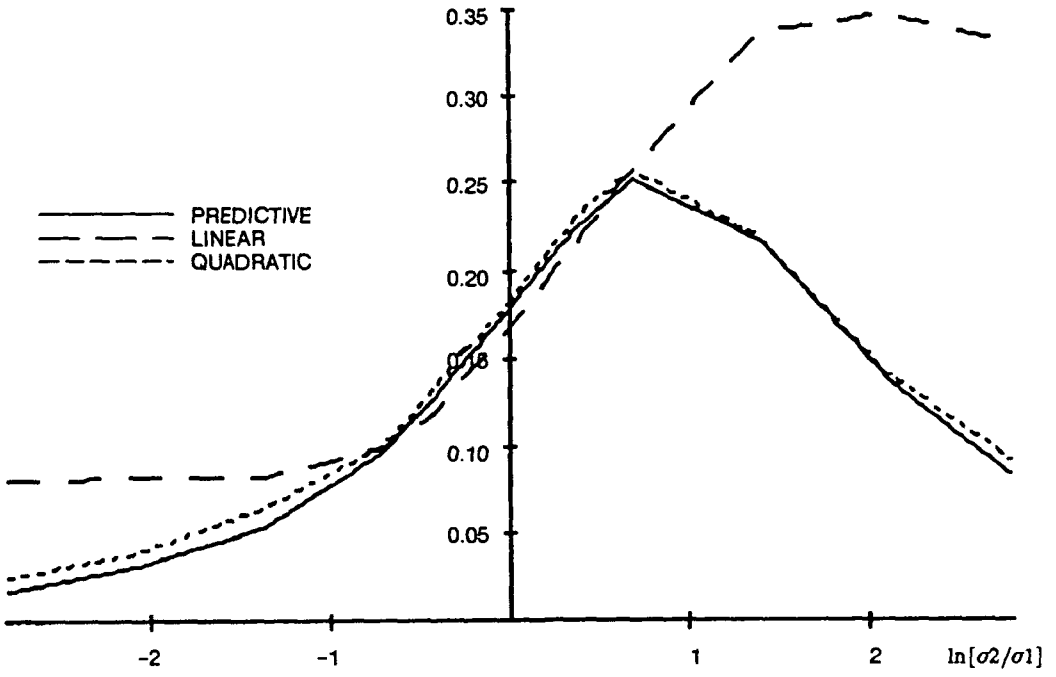

Figure 12. Average misclassification rates when distance between means is decreased $N_1 = N_2 = 6$.

Table 8. Average misclassification rates when distance between means is decreased, for various ratios of the population standard deviations $\frac{\sigma_2}{\sigma_1}$, with $N_1 = N_2 = 6$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0809	0.0822	0.0995	0.1223	0.1674	0.2193	0.2565	0.3356	0.3318
quad	0.0256	0.0604	0.1017	0.1355	0.1816	0.2330	0.2572	0.2201	0.0920
pred	0.0185	0.0541	0.0978	0.1296	0.1770	0.2279	0.2522	0.2175	0.0839

Case 9. Misclassification rate as a function of class π_2 training sample size, $\sigma_2 = \frac{1}{16}$.

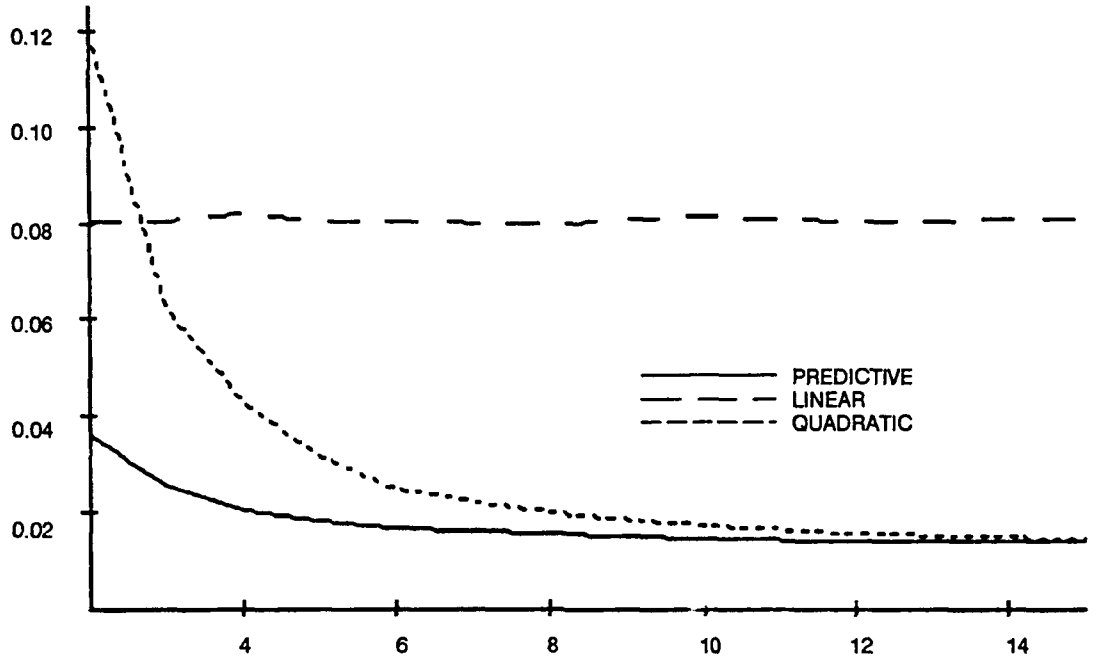

Figure 13. Misclassification rates as a function of class π_2 training sample size, with $\frac{\sigma_2}{\sigma_1} = \frac{1}{16}$.

Table 9. Misclassification rates for various ratios of N_2 , given $\frac{\sigma_2}{\sigma_1} = \frac{1}{16}$, $N_1 = 15$, $\mu_1 = 0$, $\mu_2 = 2$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.0806	0.0804	0.0821	0.0802	0.0801	0.0798	0.0815	0.0801	0.0809
quad	0.1194	0.0626	0.0427	0.0321	0.0249	0.0200	0.0172	0.0155	0.0147
pred	0.0363	0.0259	0.0206	0.0186	0.0168	0.0156	0.0148	0.0143	0.0138

As shown before and also seen here, the quadratic discriminant function seems to be unstable in the parameter estimations at low sample sizes, and the misclassification rate is worse than the linear, even though the variances are very different. The predictive function does not show this problem, but is the best discriminant function for this case.

Case 10. Misclassification rate as a function of class π_2 training sample size, $\sigma_2 = 1$.

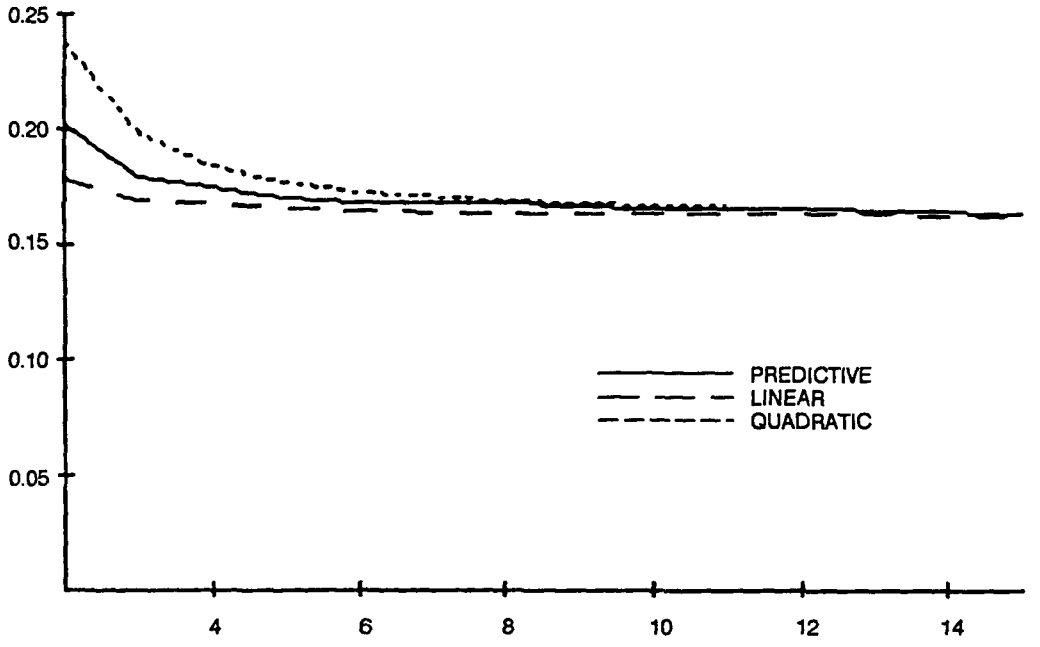

Figure 14. Misclassification rates as a function of class π_2 training sample size, with $\frac{\sigma_2}{\sigma_1} = 1$.

Table 10. Misclassification rates for various ratios of N_2 , given $\frac{\sigma_2}{\sigma_1} = 1$, $N_1 = 15$, $\mu_1 = 0$, $\mu_2 = 2$.

Disc. Func.	Ratios of Standard Deviations								
	1/16	1/4	1/2	2/3	1	1.5	2	4	16
linear	0.1782	0.1691	0.1683	0.1656	0.1647	0.1636	0.1638	0.1637	0.1625
quad	0.2377	0.1976	0.1839	0.1772	0.1724	0.1696	0.1664	0.1662	0.1635
pred	0.2017	0.1793	0.1746	0.1700	0.1685	0.1680	0.1657	0.1655	0.1634

When the variances are equal in value, the misclassification rates of the predictive and quadratic discriminant functions are marginally larger than the linear function, which should be the clear winner. Note that the predictive and quadratic functions quickly converge on the linear function in performance.

MULTIVARIATE CASE—FISHER'S IRIS DATA

An analysis was performed of the misclassification rates of the discriminant functions on a real multivariate data set to examine the effect of small sample sizes on the discriminant functions in a multivariate environment.

The data consisted of three classes of Iris flowers. This was the data set used by Fisher in analyzing the linear discriminant function. The parameters measured were petal width, petal length, sepal width, and sepal length. The data consisted of 50 measurements from each class. Figures 15, 16, and 17 are projections of the data for various features.

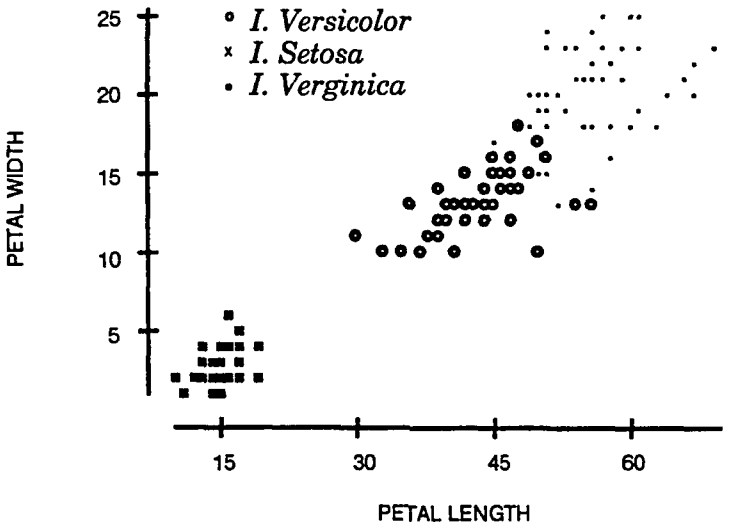

Figure 15. Fisher Iris data—petal length vs. petal width.

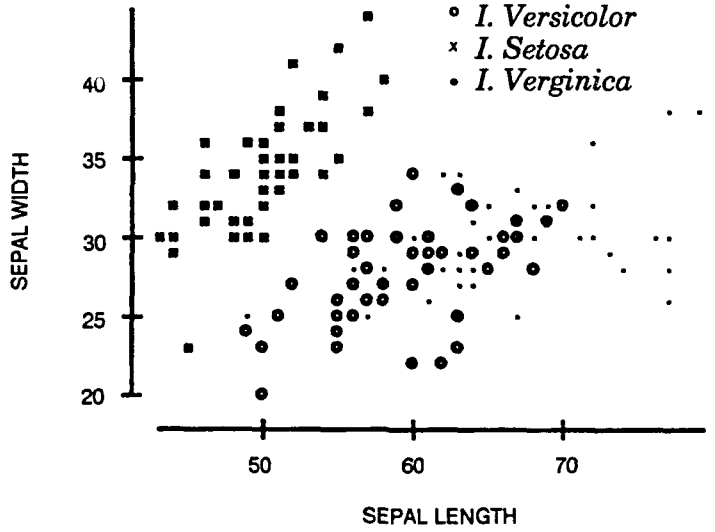

Figure 16. Fisher Iris data—sepal length vs. sepal width.

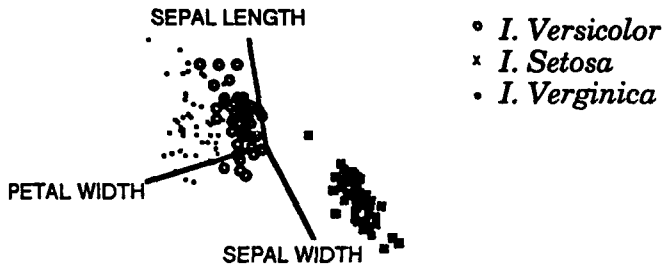

Figure 17. Fisher Iris data—sepal length vs. petal width.

Note that the *I. Setosa* is well separated from the other two plants. The misclassifications therefore will occur mostly between *I. Versicolor* and *I. Verginica*. For this reason, the simulations will examine misclassification rates when the training sample sizes of *I. Versicolor* and *I. Verginica* are modulated. Another reason for this is that the petal width of *I. Setosa* does not vary much, and generating small samples from this class quite often generates a degenerate covariance matrix. This is due to the fact that the measurements are integers, and a random set of observations might have the same value for a feature, thereby creating a singular covariance matrix. Indeed, the integer nature of all the observations preclude very small sample size comparisons (if the samples are to be independently selected).

The analysis varied the training sample sizes from the classes of data, and the probabilities of misclassification were derived through Monte Carlo simulation. The simulations were performed as follows:

1. The training sample sizes for the three classes were read from the command line. The observations were also read into the program from a file. Since it is "bad" practice to test on an observation used for training, the observations for each class were randomly divided into a training and a testing set. If an observation was used in training, it was not used to test the discriminant functions.
2. For each class, a uniform random number generator generated a set of real numbers that were converted to integers in the range of the number of observations for the class. Since duplicate numbers can be generated, the process was repeated for duplicate numbers until the members in the set were unique. These numbers were used as indexes into the array of observations, indicating the observations to be used as the training set.
3. The observations corresponding to the index values were used as the training set, and a sample mean and covariance were generated. The observations not indexed were used as the testing set.
4. The discriminant functions were tested with the test set, and the misclassification rate was computed.

5. Steps 2 through 4 were repeated for a number of iterations, and the averaged results were reported.

MISCLASSIFICATION RATES AS A FUNCTION OF TRAINING SAMPLE SIZE

To measure the misclassification rates, the training sample sizes of *I. Versicolor* and *I. Verginica* were varied and the results are shown below. Misclassification rates were measured when (1) one class' training sample size was varied, with the others' remaining constant, and (2) two classes' training samples were varied concurrently. The results are shown in figures 18 and 19, with corresponding tables 11 and 12.

Note that the projections of the covariances of the three classes of plant are quite similar, and therefore one should expect that the linear discriminant function will outperform the quadratic and predictive discriminant functions for this data set. Also, due to the discrete nature of the data, a fairly large set (> 10) of class π_1 (*I. Setosa*) training samples was required during the simulations. This is due to the fact that the petal length of the *I. Setosa* is predominantly the value 2, and invariably a small random sample of the class would cause a singular covariance matrix (all the training samples would have a petal length of 2). The large set of *I. Setosa* helps the linear discriminant function stabilize the pooled covariance matrix and thereby improves the linear function's performance.

Case 1. *I. Versicolor*'s training sample size varied, other classes' training sizes constant.

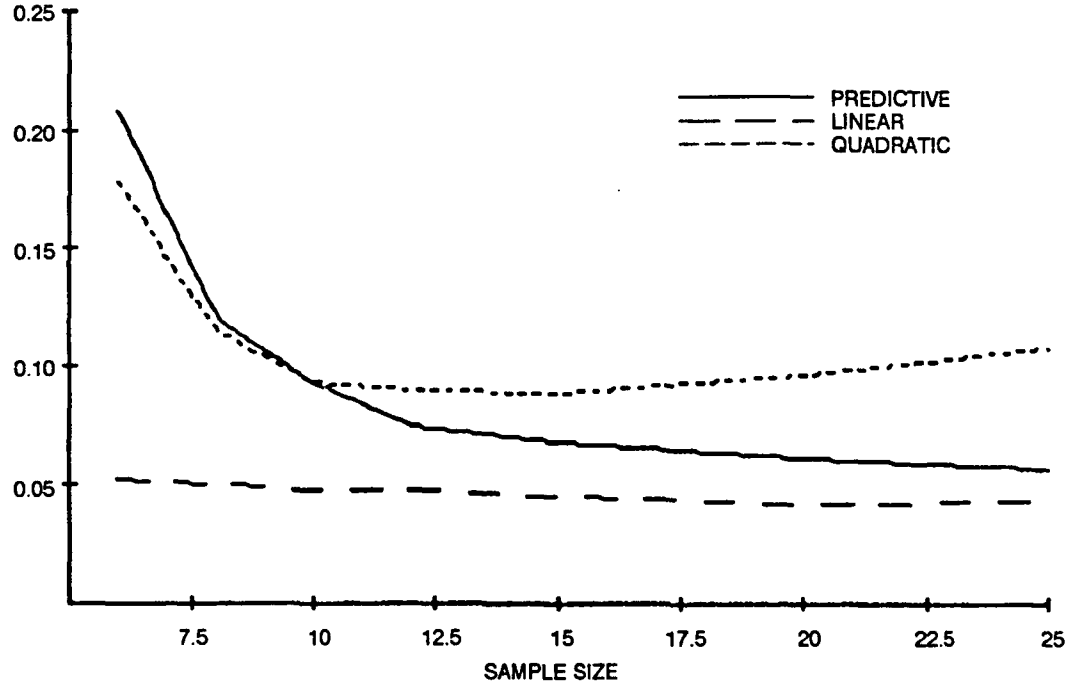

Figure 18. Misclassification rates for various sample sizes of *I. Versicolor*.

Table 11. Misclassification rates for various values of N_2, given $N_1 = 20$, $N_3 = 15$.

Disc. Func.	Values of N_2						
N_2	6	8	10	12	15	20	25
linear	0.0519	0.0497	0.0483	0.0482	0.0451	0.0424	0.0429
quad	0.1783	0.1145	0.0937	0.0898	0.0885	0.0968	0.1080
pred	0.2074	0.1213	0.0937	0.0752	0.0681	0.0609	0.0571

In this case, the predictive function shows more instability at lower sample sizes than the quadratic function, but at higher sample sizes converges quickly to the performance of the linear discriminant function. Again, it is expected that the linear function will perform best because of the similarity of the covariance matrices of the classes. Notice that the quadratic function is not converging, but is diverging at the higher sample sizes. Whether this is a true divergence or just an anomaly due to the variation in sampling size is not an issue, but the quadratic function is not converging to the performance of the linear discriminant function for these sample sizes.

Case 2. *I. Versicolor* and *I. Verginica* training sample sizes varied together, *I. Setosa* training size constant.

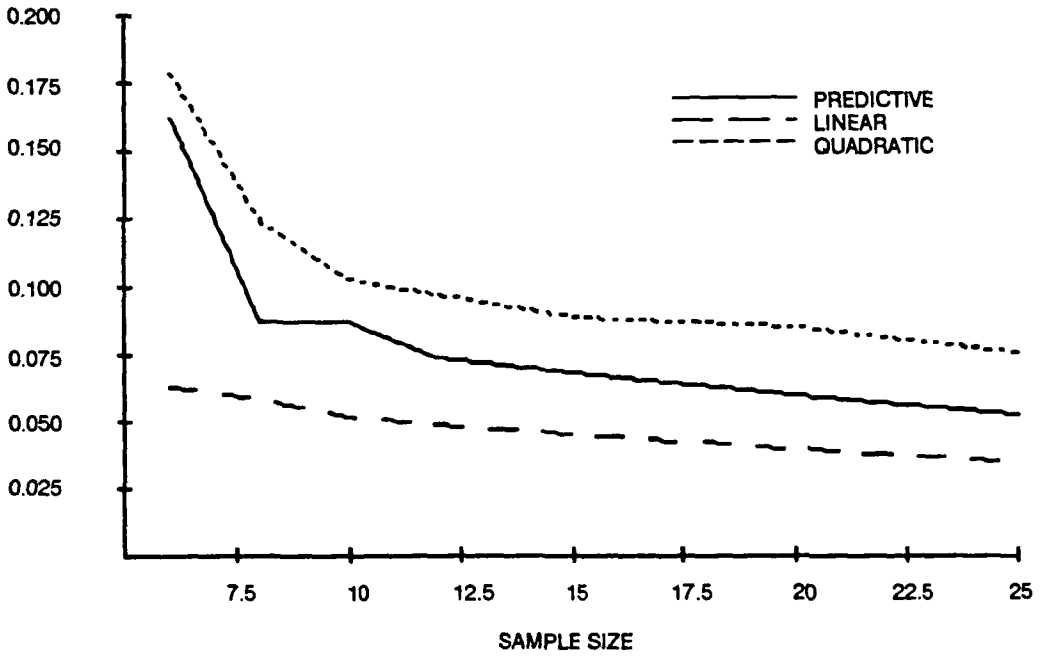

Figure 19. Misclassification rates for various equal sample sizes for *I. Versicolor* and *I. Verginica*.

Table 12. Misclassification rates for various values of $N_2 = N_3$, given $N_1 = 20$.

Disc. Func.	Value of N_2 and N_3						
N_2, N_3	6	8	10	12	15	20	25
linear	0.0624	0.0581	0.0516	0.0485	0.0451	0.0399	0.0350
quad	0.1783	0.1241	0.1025	0.0973	0.0885	0.0848	0.0758
pred	0.1618	0.1165	0.0871	0.0739	0.0681	0.0602	0.0524

Note that this test gives the quadratic discriminant function a better chance to perform well, since one of the underlying assumptions of the quadratic function is that of equal training samples from each class. The two classes which overlap in the feature space (*I. Versicolor* and *I. Verginica*) are the two classes with an equal number of training samples. Note, however, that the quadratic discriminant function is still outperformed by the predictive discriminant function.

CONCLUSION

UNIVARIATE CONCLUSIONS

Predictive Discriminant Function Versus Quadratic Discriminant Function

The quadratic discriminant function is asymptotically predictive—as sample sizes go up, the quadratic function's misclassification rates approach those of the predictive discriminant function. At low sample sizes, quadratic function shows instability in variance estimates for well-separated populations, and its performance can be worse than even that of the linear discriminant function for small sample sizes and unequal variances. A very important point is that, in every case, and in every univariate simulation made, the likelihood-based quadratic function has been inferior (i.e., higher misclassification rates) to the predictive discriminant function.

Predictive Discriminant Function Versus Linear Discriminant Function

The linear discriminant function outperforms the predictive discriminant function when covariances are near-equal, because the assumption of equal variance allows the use of pooled variance for the linear discriminant. Linear performance is poor when the underlying population variances are not close in value. Note that even when the variances are equal, the predictive discriminant function performs nearly as well as the linear discriminant function, especially when compared to the performance of the quadratic discriminant function.

MULTIVARIATE CONCLUSIONS

Predictive Discriminant Function Versus Quadratic Discriminant Function

In the multivariate case, the predictive discriminant function displays the interesting ability of instability when the sample sizes are small. However, the performance of the quadratic discriminant function is shown to be equally unstable, and the quadratic function misclassification rate decreases more slowly as sample size increases. This instability is understandable when one understands that both the quadratic and predictive discriminant functions are attempting to estimate a four-dimensional covariance matrix with a sample size of as few as six observations. It is therefore understandable that the error rates for these discriminant functions are quite large with six points. Note that the predictive discriminant function falls quickly as a function of the training sample size n, with $n = 2$ or 3 times the dimension of the observation vector $\mathbf{X}$ for a fairly stable estimate of the covariance matrix.

Predictive Discriminant Function Versus Linear Discriminant Function

Since the data in the multivariate case appear to have similar covariance structures, it was expected that the linear discriminant function would outperform the predictive discriminant function for the Fisher Iris data. This has been shown to be true. The linear discriminant function is able to use the similarity of the covariance matrices to stabilize its estimate of the covariance structure of the classes.

FINAL STATEMENT

To conclude, the predictive and linear discriminant functions generally are (1) and (2) in performance. The linear function performs better when variances are similar, while it performs poorly if the variances of the distributions vary widely. The predictive discriminant function shows better performance than the quadratic function in almost all situations, since it is taking into account more information (i.e., sample sizes). Therefore, in case of classes with small training sample sizes, the predictive discriminant function is preferable to the quadratic discriminant function in minimizing misclassification rates.

This work has shown that, for small training set classifications, the predictive discriminant function should not be neglected. The predictive discriminant function is versatile in its application, minimizes the number of assumptions made, and performs reasonably well over the range of cases tested.

REFERENCES

Anderson, T. W., 1984. *An Introduction to Multivariate Statistical Analysis* (2nd ed.), John Wiley & Sons, New York.

Duda, R. O., and P. E. Hart. 1973. *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York.

Enis, P., and S. Geisser. 1974. Optimal Predictive Linear Discriminants, *The Annals of Statistics*, 1974, Vol. 2, No. 2, pp. 403–410.

Geisser, S. 1964. "Posterior Odds for Multivariate Normal Classifications," *Journal of the Royal Statistical Society*, Series B, Vol. 26, 1964, pp. 69–76.

Kendall, M., A. Stuart, and J. K. Ord. 1987. *Kendall's Advanced Theory of Statistics*, Volume 3, Oxford University Press.

Marks, S., and O. J. Dunn. 1974. "Discriminant Functions When Covariance Matrices are Unequal," *Journal of the American Statistical Association*, June 74, Vol. 69, No. 246, pp. 555–559.

Press, S. J. 1982. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, Krieger.

Raudys, S. J., and A. K. Jain. 1991. "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 3, 1991, pp. 252–264.

Wahl, P. W., and R. A. Kronmal. 1977. "Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate," *Biometrics 33*, pp. 479–484.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1992	Final: Oct 91 – Jan 92

4. TITLE AND SUBTITLE

MISCLASSIFICATION RATES OF LIKELIHOOD AND PREDICTIVE DISCRIMINANT FUNCTIONS FOR SMALL SAMPLES

5. FUNDING NUMBERS

0305885G
CD38
1CCD38D0

6. AUTHOR(S)

Don Waagen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Command, Control and Ocean Surveillance Center (NCCOSC)
RDT&E Division (NRaD)
San Diego, CA 92152–5000

8. PERFORMING ORGANIZATION REPORT NUMBER

NRaD-TD 2277

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Security Agency
Fort George G. Meade, MD 20755

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Likelihood and predictive discriminant function misclassification rates are compared when training sets are small (i.e., less than 20). The theoretical foundations for the linear, quadratic, and predictive discriminant functions are described. Simulations are used to compare the classification capability of each discriminant function while varying the variance of the underlying distributions. A multivariate case is also analyzed.

14. SUBJECT TERMS

classification
statistics
statistical pattern recognition
Discriminant Analysis

15. NUMBER OF PAGES

40

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT

21a. NAME OF RESPONSIBLE INDIVIDUAL	21b. TELEPHONE *(Include Area Code)*	21c. OFFICE SYMBOL
Don Waagen	(619) 553-6023	Code 422

INITIAL DISTRIBUTION

Code 0012	Patent Counsel	(1)
Code 144	V. Ware	(1)
Code 40	R. C. Kolb	(1)
Code 42	J. Salzmann,Jr.	(1)
Code 422	D. K. Porter	(1)
Code 422	D. Marchette	(1)
Code 422	D. Waagen	(5)
Code 952B	J. Puleo	(1)
Code 961	Archive/Stock	(6)
Code 964B	Library	(2)

Defense Technical Information Center
Alexandria, VA 22304-6145 (4)

NCCOSC Washington Liaison Office
Washington, DC 20363-5100

Center for Naval Analyses
Alexandria, VA 22302-0268

Navy Acquisition, Research & Development Information Center (NARDIC)
Alexandria, VA 22333

National Security Agency
Fort George G. Meade, MD 20755 (2)